Attention-based deep learning model for facial expression recognition

by

Alimzhan Kairzhanov

Submitted to the Computer Science
in partial fulfillment of the requirements for the degree of Master

of Science in Data Science

at the NAZARBAYEV

UNIVERSITY

July 2022

Author . . . .

Alimzhan Kairzhanov

Computer Science

June 10, 2022

Certified by. . . . .

Anh Tu Nguyen

Assistant Professor, School of Engineering and Digital Sciences

Thesis Supervisor

Certified by. . . . .

Min-Ho Lee

Assistant Professor, School of Engineering and Digital Sciences

Thesis Supervisor

Accepted by . . . .

Vassilios D.Tourassis

Dean, School of Engineering and Digital Sciences

Attention-based deep learning model for facial expression recognition

by

Alimzhan Kairzhanov

Submitted to the Computer Science
on June 10, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Facial expression recognition is an active area of research in computer vision and deep learning, which has become popular in recent decades. The results of these studies are used in psychology, behavioral science and computer-human interaction. Emotion recognition is a very difficult task, since it is necessary to overcome such difficulties as the presence of a large number of images, head rotation, lighting conditions, partial face closure (glasses, mask, hand, etc.) In this regard, in this practical study, we use different models of Vision Transformer (ViT) to improve the accuracy of classification on publicly available datasets of CK+ and JAFFE. The results obtained show that we have achieved excellent accuracy values compared to state-of-the-art works using a fewer computational resource to train.

Keywords— facial expression recognition, Vision Transformer, attention mechanism, image classification

Thesis Supervisor: Anh Tu Nguyen
Title: Assistant Professor, School of Engineering and Digital Sciences

Thesis Supervisor: Min-Ho Lee
Title: Assistant Professor, School of Engineering and Digital Sciences

Acknowledgments

I want to express huge gratitude to my adviser Anh Tu Nguyen, who provided incredible support and gave me a lot of valuable resources. He really helped me to go this long way in various fields .

Contents

Chapter 1

Introduction

1.1 Overview and motivation

Face plays an important role in people's communication. It is a reflection of a person's personality, his thoughts and emotions. Human communication can be divided into two parts: verbal and nonverbal. According to a psychological study conducted by Mehrabian [31], the nonverbal part is the most informative in social interaction. So, the verbal part is about 7% of all information, the vocal part is 34%, and the facial expression is 55%. For this reason, a person is the object of research in many fields of science, such as psychology, behavioral research, computer-human interaction, medicine.

In the last century, Ekman and Friesen [16] identified six fundamental emotions based on an interracial study that confirms the fact that people's emotions manifest themselves equally regardless of culture. These fundamental emotions are anger, sadness, surprise, happiness, disgust and fear. In addition to the six main ones, researchers in this field consider the seventh facial expression – neutral.

1.2 Problem statements

The modern development of artificial intelligence technologies in the field of image classification is aimed at automatic identification of images. Computers have learned to "understand" a person's mood and react accordingly. And since this is not an easy area of research due to the complexity of the nature of emotions and precise facial features, there is still room for improvements in recognition accuracy in this area.

With the spread of machine learning, especially deep learning, researchers have achieved significant results in emotion recognition in the last quarter of a century. Prior to the widespread use of deep learning, traditional emotion recognition methods used shallow learning and handcrafted features (non-negative matrix factorization (NMF) [6], Local Binary Patterns (LBP) [4], [5], Histograms of Oriented Gradients (HOGs) [3] and sparse representation [7]). The growth of deep learning-based approaches has resulted in current indicators (for example [8], [9], [10], [11]).

1.3 Aims and objectives

Recently, impressive works on facial expression recognition have been published. However, they use traditional convolutional networks, and deep learning has rarely been transformed. When recognizing emotions, only certain parts of the face, such as eyebrows, eyes and mouth, carry the most information. While hair and ears are not involved in the expression of emotions. Therefore, modern models should pay attention only to informative sections. Not so long ago, attention models were successfully applied to FER to study significant regions. Li et al. [27] proposed a CNN with patch-gating that combines attention at the pathway level for expression recognition with occlusion.In expansion, attention models have been effectively connected to FER to ponder noteworthy regions. Essentially to [27], a few strategies such as [46], [19], [24], attention-like instruments were utilized to center on the foremost unmistakable highlights to move forward the precision of the FER. In the work [45], a Transformer based on the mechanism of attention was presented. The new transformer architecture [45] has led to a big leap forward in the possibilities of sequential modeling in NLP problems. The great success of transformers in NLP has aroused particular interest from the vision community in understanding whether transformers can be a strong competitor to the dominant architectures based on convolutional neural networks (CNN) in vision tasks such as ResNet [20] and EfficientNet [43]. In this work, different models of ViT are used, the experimental results of which surpass traditional convolutional networks in terms of accuracy.

The aim of this study is to increase the level of accuracy of emotion recognition for a more accurate classification when combining the mechanism of attention and the use of a ViT model. To begin with, we will output accuracy indicators on convolutional neural networks using pre-trained VGG, ResNet models. Next, using the attention mechanism, we will reduce the concentration area of the model to only the necessary parts of the face. The task of emotion recognition is currently quite relevant in various fields of activity, such as sociology, the gaming industry, robotics and human-computer interaction.

1.4 Key contributions

The main contributions of our work are as follows:

1. The empirical examination of several ViT models based on attention mechanisms for FER.

2. Experimental results on publicly available datasets, such as JAFFE and CK+48, show that current ViT model and its variants demonstrate a great potential in achieving the state-of-the-art performance.

This work is organized as follows. Section 2 provides an overview of previous work in this area. Section 3 will be devoted to the proposed structure and architecture of the model. After that, in section 4, we will present the experimental results, describe the datasets used in this article and compare them with modern works. In conclusion, we will conclude the article in section 5 and consider the areas of further research.

Chapter 2

Related works

2.1 Facial expression classification

To date, many facial expression recognition systems work automatically, classifying by one of the 7 basic emotions: anger, sadness, surprise, happiness, disgust, neutral and fear. Among the variety of ways of encoding emotions, the most popular is the "Facial Action Coding System" (FACS), developed by Paul Ekman and Wallace Friesen [17]. The scope of this standard classification of facial expressions varies from medicine to computer animation.

In traditional methods, the stages of classification and extraction of objects are independent. For example, Haar Cascade is an object detection algorithm used to identify faces in an image or a real time video. The algorithm uses edge or line detection features proposed by Viola and Jones in their research paper "Rapid Object Detection using a Boosted Cascade of Simple Features" published in 2001. The first contribution to the research was the introduction of the haar features. These features on the image makes it easy to find out the edges or the lines in the image, or to pick areas where there is a sudden change in the intensities of the pixels. The main advantages of this method are as follows: Haar-like features are more robust to illumination changes than color histogram; The feature-based system operates much faster than a pixel-based system; The Integral Image allows the sum of pixel responses within a given sub-rectangle of an image to be computed quickly; Only several accesses to the integral image are required to extract a Haar-like feature response; Allows real time detection. Along with the advantages, there is also the main drawback that Haar-like features are not invariant over rotation. This means that any object that rotates is sensitive to angle changes will be difficult to solve using standard Haar-like features.

The local binary pattern (LBP) operator is an image operator which transforms an image into an array or image of integer labels describing small-scale appearance (textures) of the image. These labels directly or their statistics are used for further analysis. The main advantages of this method are: High discriminative power; Computational simplicity; Invariance to grayscale changes and good performance. The disadvantage is also not invariant to rotations and the size of the features increases exponentially with the number of neighbours which leads to an increase of computational complexity in terms of time and space.

In contrast, deep networks perform FER in an end-to-end way. A loss layer is added to the end of the network to regulate the backpropagation error; after that, the network outputs the probability of predicting each sample. To minimize the cross entropy between the estimated class probabilities and the truth distribution, the softmax loss function is most often used. In [44], the authors demonstrated the advantage of using a linear support vector machine (SVM) for end-to-end learning. Instead of cross entropy, it minimizes losses based on margin. In the same way, by replacing the loss of softmax with the adaptation of deep neural forests (NFs), the authors of the study [11] achieved visible results.

The use of a deep neural network as a complement to the end-to-end learning method is used as a feature extraction tool. Further, additional independent classifiers are applied to the extracted representations, such as a random forest or a support vector machine [13, 39].

2.2 Convolutional neural networks with Attention

In recent years, researchers have been actively using convolutional neural networks to detect objects and classify images. Convolutional network layers automatically extract representations from input images. At the initial layers, the basic image properties are extracted, such as the edges of objects, shapes, and various colors when working with color images. At the later layers, specific properties are extracted depending on the data set and the task at hand. At the final stages, fully connected layers are connected, which process the data of the previous layers and give the result inherent to one of the classes [35]. The most popular methods of emotion classification are Haar features [50], local binary patterns (LBP) [38] and histogram of oriented gradients (HOG) [8]. On small data sets created in the laboratory, they show good results, however, with an increase in the amount of input data, changes in image creation conditions (such as illumination, face pose, incomplete face image, etc.), recognition accuracy indicators decrease.

Previously, deep convolutional neural networks turned out to be the most popular for image classification [41]. The bottom line of transfer learning method is that the properties and skills extracted from the previous task can be applied in a new task [33]. The main goal is to apply knowledge to the target area. One of such ways of using the pre-trained models on the ImageNet dataset [41] is applied in replacing the last fully connected layers with layers aimed at the current task. These characteristics are used to train classifiers such as Softmax and Long short-term memory (LSTM). The main part of the pre-trained model remains unchanged. Next, there are two ways to adjust the weights. The first is to train the model from scratch: with the setting of random values of weights and further adjustment. This takes much longer, because the number of parameters being trained increases significantly. The second method is to use the frozen weights of the pre-trained model and adjust only the weights on the last modified layers. This method takes less time to set up and requires less computing resources.

The author A. Ravi in his work [35] classifies 4 methods of using transfer learning for a target task, depending on the size and similarity with the original data set. So, with a small set with great similarity to the original, there is a high probability of retraining the model. With a large set with a similar to the original, it is possible to achieve the desired results. The third option is obtained with a small data set with a big difference from the original. And finally, the fourth one consists of a huge amount of input data and is very different from the original dataset. In the latter case, a convolutional neural network can be trained from arbitrary weights, but most researchers use established weights [22, 40, 36].

Attention has been widely employed to improve feature representations in a variety of ways. For example, SENet [21] employs channel-attention, CBAM [51] adds spatial attention, and ECANet [47] suggests an efficient channel attention to improve SENet further. Combining CNNs with various forms of self-attention has also piqued curiosity [4, 42, 56, 34]. To replace the convolutional layer, SASA [34] and SAN [56] use a local-attention layer. Prior approaches, despite promising results, limited the scope of focus to the immediate region due to its complexity. LambdaNetwork [4] has introduced an efficient global attention model to model both content and position-based interactions in picture classification models, significantly improving the speed-accuracy tradeoff. In the final three bottleneck blocks of a ResNet, BoTNet [42] replaces spatial convolutions with global self-attention, resulting in models that perform well on the ImageNet benchmark for image categorization. Unlike these techniques, which combine convolution and self-attention, our work is based on a pure self-attention network, such as ViT [14], which has lately shown considerable promise in a variety of vision applications.

2.3 Vision Transformer

At the same time, in this work, along with transfer learning, a ViT is used. The first application of a ViT for image classification is shown in [14]. The work of Vaswani et al. [45] was taken as a basis, where the authors for the first time introduce the concept of Transformers with application in natural language processing. The model was pre-trained on the ImageNet database [12] and its indicators are superior to modern models. Huge datasets (exceeding 100 million images) are required to train the model and adjust the weights. In this regard, it can be concluded that the ViT refers to models with high data consumption.

In instance, ViT [14] is the first transformer-based image categorization approach to match or even beat CNNs. Several researchers have attempted to invest Transformers in computer vision tasks such as object identification [3], posture estimation [53], high-resolution image synthesis [18], video instance segmentation [49], trajectory prediction [5], and so on, inspired by the popularity of Transformers. When completely trained on large-scale datasets, transformer-based algorithms have shown greater performance over CNN-based methods. The first work to apply a vanilla Transformer on photos with minor changes was ViT [14]. When trained on ImageNet [12], ViT had poorer accuracy than ResNet, according to [14]. Because Transformers require a considerable amount of data to generalize effectively on computer vision tasks, ViT was first trained on big datasets and then finetuned for downstream applications. Transformers have used the feature pyramid structure seen in CNNs. For pixel-level dense prediction, Wang et al. [48] suggested Pyramid Vision Transformer (PVT), which can operate as the feature extraction backbone without convolutions. Several publications [9], [52], [29] advocated blending convolutional layers into Transformers, which enhanced the performance of pure Transformers even more. We propose to use Transformers directly for FER, inspired by the vanilla Transformer and these great Transformer-variants. As far as we know, no effort has attempted to capture the correlations between deep characteristics in order to recognize face expressions. We use Transformers to simulate the self-attention mechanism's lengthy dependencies between input sequences. In the event of occlusions or alternative postures for FER, such self-attention allows the model to disregard the information-deficient regions and detect the expressions from a global perspective.

Chapter 3

Methodology

In this chapter, we will present the proposed method in four sections: architecture overview, data preprocessing and augmentation, a CNN-based approach, and an attention mechanism for the FER.

3.1 Architecture Overview

The proposed solution consists of two components: a pre-trained model and a convolutional attention network for classifying emotions. The standard procedure for determining facial expressions is shown by the flowcharts in the figure 3-1. The first stage of the input image is the preprocessing stage, which includes face alignment, data augmentation and face normalization. The second step is to extract features from the image. Deep learning tries to capture high-level abstractions via hierarchical structures comprised of numerous nonlinear transformations and representations [25]. After extracting the features, the final stage is the classification of the image according to one of the basic emotion categories.

3.2 Data preprocessing and augmentation

The preprocessing stage is necessary to bring the input data to a single form. It's no secret that many images obtained in a natural environment contain irrelevant

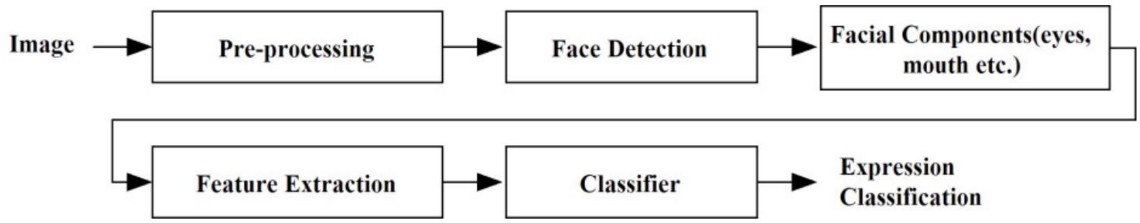

Figure 3-1: The general pipeline of facial expression recognition systems.

information, such as the rotation of the head pose, different lighting levels and background. For this reason, preprocessing is a standard technique in the field of emotion recognition.

To ensure generalizability, deep neural networks, and even more so a ViT, require a large amount of training data. The number of images in many datasets is not enough for training. In this regard, data generation technique is applied. In this work, the operations of horizontal and vertical flipping, rotation and transformation at the pixel level are applied. The combination of various operations creates a larger amount of data [55], [26].

3.3 CNN-based approaches

In a classical convolutional neural network, the neurons of the previous layer are connected to the next one. Each compound forms certain weights. The architecture of deep learning is an array (set) of weights. When training a model from scratch, random values are set to weights and recognition accuracy starts with insignificant numbers. To save time and take into account the limited computing capabilities of the hardware, transfer learning is used in this work. The transfer learning technique is a popular method of building models in a timesaving way where learning starts from patterns that have already been learned [32, 54]. The repurposing of pre-trained models avoid straining from the sketch that requires a lot of data and leverages the huge computational efforts. In other words, transfer learning reuses the knowledge through pre-trained models [37] that have been trained on a large benchmark dataset for a similar kind of problem.

If there is not enough data training the model gives little accuracy. This fact is explained by the lack of strong regularization.

The opposite situation occurs with large databases. Experiments show that retraining on significant sets overshadows inductive bias. Also, the results are impressive when shifting to tasks with fewer outputs. In our case, there are 7 classes of emotion expression.

The approach based on convolutional neural networks uses such pre-trained models as VGG19 and ResNet50. We use the SVM classifier to achieve our goals. The structure of the VGG19 model consists of 19 convolutional and fully connected layers divided into 5 groups. The output of each of them is used to evaluate the best features. As with any linear classifier, this model has updatable weights and biases. The total number of trained parameters exceeds 20 million. By default, the input parameters of the image are 48x48 RGB, but we change the size to 224x224.

In turn, ResNet50 consists of 50 layers. We replaced the output layers of the original model with an alignment layer and added 3 fully connected layers. The last softmax layer contains 7 output classes. Most of the pre-trained model was frozen, while the remaining part was subjected to training. We used Adam as an optimizer, with a learning coefficient of 0.0005 and a bucket size of 10, the number of epochs was 50.

As mentioned earlier, transfer learning is notable for using skills and knowledge from previous tasks to apply to new tasks [33]. With the freezing of most layers, fewer parameters are retrained, which allows us to spend less time on training and save computing resources.

3.4 Attention mechanisms for FER

It is a well-known fact that not all parts of the face take the same part in the formation of emotions. Potential areas of emotion formation can be called special areas, such as the mouth, eyebrows, eyes. Based on this conclusion, we have built a self-attention model that pays attention only to the important regions of the face.

Attention mechanisms are increasingly used to model sequences, since they do not depend on the distances in the input and output sequences [2], [23]. The combination with a recurrent network remains a priority for this mechanism. The main advantage of Transformers over a convolutional neural network is excellent results combined with significantly less computational resources for training.

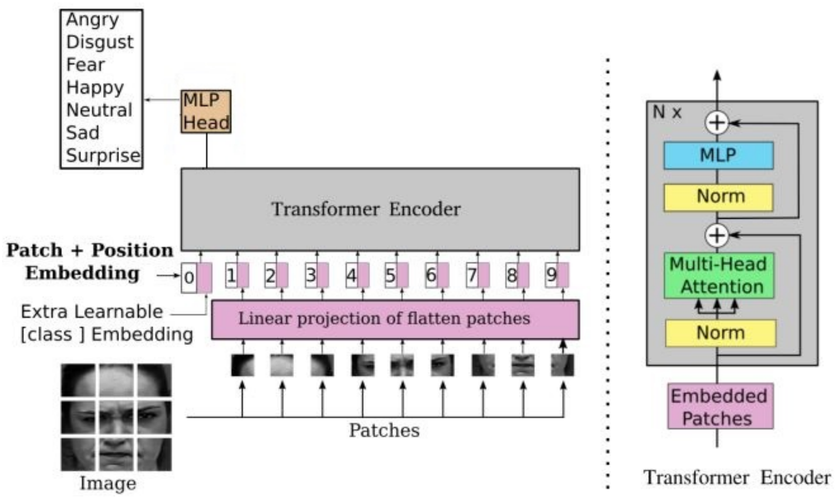

Figure 3-2: Vision Transformer model overview.

The ViT is a model for image classification that employs a Transformer-like architecture over patches of the image. This includes the use of Multi-Head Attention, Scaled Dot-Product Attention and other architectural features seen in the Transformer architecture traditionally used for Natural Language Processing.

In this work we divide a picture into fixed-size patches, linearly embed each, add position embeddings, and feed the resultant vector sequence to a typical Transformer encoder. To conduct classification, we employ the conventional method of inserting an extra learnable "classification token" into the sequence. The Transformer encoder (Figure 3-2) [14] consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks. Layernorm (LN) is applied before every block, and residual connections after every block. The MLP contains two layers with a GELU non-linearity. Encoder processes the input information, searches for important parts and creates attachments for each patch of the image based on the correspondence of other patches in the whole image.

In deep learning, attention may be widely viewed as a vector of importance weights: to forecast or infer one element, such as a pixel in an image or a word in a phrase, we estimate how strongly it is connected with other elements using the attention vector and use the sum of their values weighted by the attention vector as an approximation of the target.

The major component in the transformer is the unit of *multi-head self-attention mechanism.* The transformer views the encoded representation of the input as a set of **key-value** pairs, (**K, V**), both of dimension n (input sequence length); in the context of NMT, both the keys and values are the encoder hidden states. In the decoder, the previous output is compressed into a **query** (**Q** of dimension m) and the next output is produced by mapping this query and the set of keys and values.

The transformer adopts the scaled dot-product attention: the output is a weighted sum of the values, where the weight assigned to each value is determined by the dot-product of the query with all the keys:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V}$$

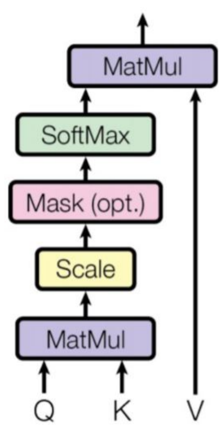

Figure 3-3: Scaled Dot-Product Attention [45].

Rather than only computing the attention once, the multi-head mechanism runs through the scaled dot-product attention multiple times in parallel. The independent attention outputs are simply concatenated and linearly transformed into the expected dimensions. According to the paper [45], *"multi-head attention allows the model to* *jointly attend to information from different representation* ***subspaces*** *at different positions. With a single attention head, averaging inhibits this."*

$$\begin{aligned}\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= [\mathrm{head}_1;\ldots;\mathrm{head}_h]\mathbf{W}^O\\ \text{where } \mathrm{head}_i &= \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V)\end{aligned}$$

Above **W** are all learnable parameter matrices.

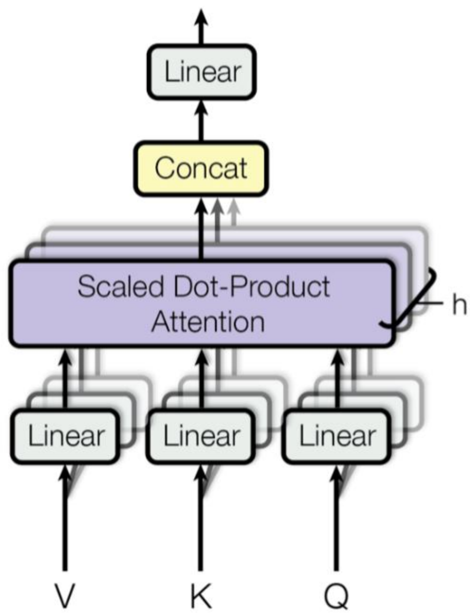

Figure 3-4: Multi-Head Attention consists of several attention layers running in parallel [45].

3.5 ConViT

Building on the insight of [10], we use the ConVit, a variant of the ViT [14] obtained by replacing some of the SA layers by a new type of layer which we call *gated positional self-attention* (GPSA) layers. The core idea is to enforce the "informed" convolutional configuration in the GPSA layers at initialization, then let them decide whether to stay convolutional or not.

The ConViT research paper [15] also builds on top of this insight and replaces the first 10 self-attention layers of the ViT with gated positional self-attention (GPSA) layers - which upon initialization act as convolutional layers and based on a gating

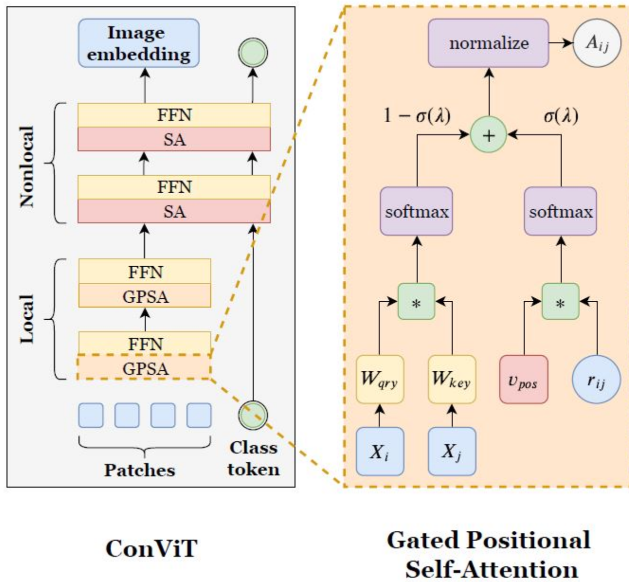

Figure 3-5: Architecture of the ConViT [15].

parameter can convert to self-attention layers. Doing so makes the earlier part of the network upon initialization behave as a convolutional neural network with the option to turn into a fully self-attention-based network based on the gating parameter which is learned via model training.

As part of this work, we are going to be looking into the ConViT architecture in detail and also look at how the GPSA layers are different from self-attention (SA) layers. Recently, the success of ViT demonstrates that the transformer architecture can be extremely powerful in data-plentiful regimes (when there is huge amounts of data available). The ViT architecture requires pretraining on huge amounts of data - JFT-300M or ImageNet-21k datasets. This is not always possible as practitioners might have sufficient hardware required to perform this pretraining. On the other hand, we know that convolutional models such as EfficientNets, can have a strong performance on fewer data as well. For example, EfficientNet-B7 was able to achieve 84.7% top-1 accuracy without any external pretraining. The practitioner is therefore confronted with a dilemma between using a convolutional model, which has a higher performance floor but a lower performance ceiling, or a self-attention-based model, which has a lower performance floor but a higher ceiling.

3.6 CrossViT

Cross-Attention Multi-Scale Vision Transformer (CrossViT) model is primarily composed of K multiscale transformer encoders where each encoder consists of two branches: (1) L-Branch: a large (primary) branch that utilizes coarse-grained patch size (P_l) with more transformer encoders and wider embedding dimensions, (2) SBranch: a small (complementary) branch that operates at fine-grained patch size (P_s) with fewer encoders and smaller embedding dimensions. Both branches are fused together L times and the CLS tokens of the two branches at the end are used for prediction.

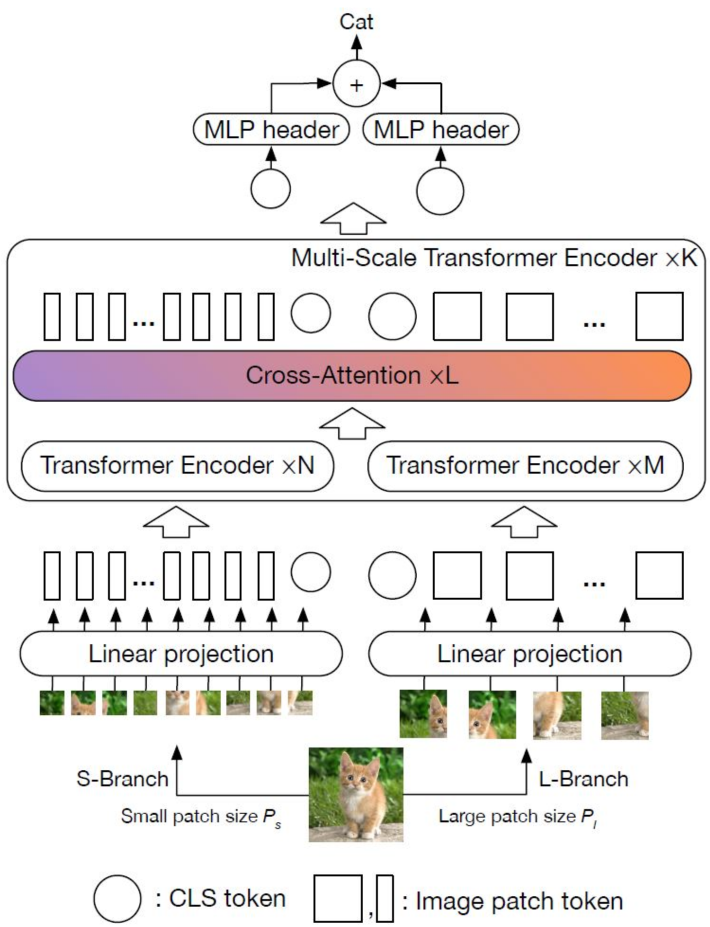

Figure 3-6: Architecture of the CrossViT [7].

CrossViT architecture (Figure 3-6) consists of a stack of K multi-scale transformer encoders. Each multi-scale transformer encoder uses two different branches to process image tokens of different sizes (P_s and P_l, $P_s < P_l$) and fuse the tokens at the end by an efficient module based on cross attention of the CLS tokens. Design includes different numbers of regular transformer encoders in the two branches (i.e. N and M) to balance computational costs.

Chapter 4

Experiments and Comparison

In this chapter, we will analyze the results obtained on some publicly available datasets and show the effectiveness of ViT over pre-trained models. First we will give a brief description of each data set used in this work. Next, we formulate the architecture of the model and the applied parameters. Subsequently, let's compare the results obtained using ViT with the results of a pre-trained convolutional neural network model.

4.1 FER datasets

JAFFE: The Japanese Female Facial Expression (JAFFE) Dataset is one of the very first datasets. It contains 213 images of 10 Japanese models who agreed to take part in the experiment and expressed 7 basic emotions. The size of the images is 256x256 pixels, the image format is .tiff. Some examples of images from the JAFFE dataset are shown in Figure 4-1.

CK+: 123 models of various ages and genders participated in The Extended Cohn-Kanade (CK+) dataset. Although the original dataset contains a sequence of images from neutral to peak emotion, we used a modified version of this dataset, which uses the last 3 peak emotions of each expression. The result was a set consisting of 981 images. We assess our method's generalization capacity using the overall sample accuracy and confusion matrices. 6 emotions, with the exception of the neutral one,

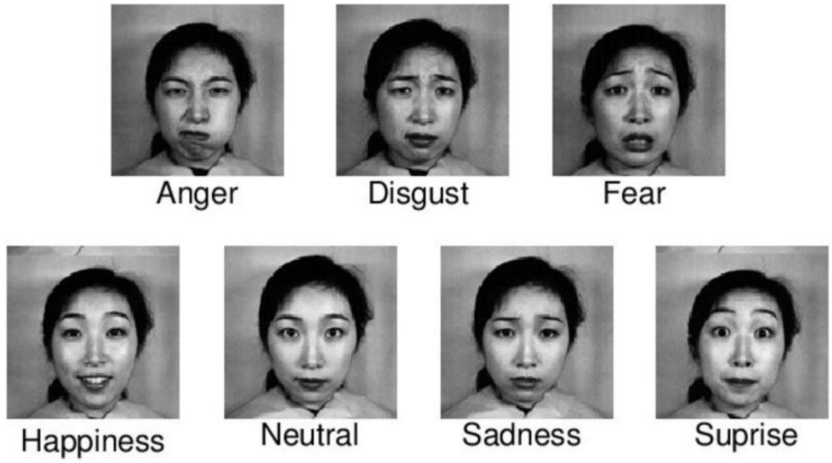

Figure 4-1: Emotion samples from JAFFE database

from the CK+48 dataset are shown in Figure 4-2.

Figure 4-2: Emotion samples from CK+48 database

4.2 Architecture and training parameters

In experiments with the ViT, we used the "timm" package model. The size of the input data is set to 224x224. As described earlier, in order to increase the number of images for training, various operations were used to increase the amount of data. The various models used in this work have a different number of trainable parameters, which affects the learning rate of the model.

4.3 FER accuracy with different deep models

We will now give the results of the suggested model on the aforementioned datasets. In each scenario, we train the model on a portion of the dataset, validate it on the

№	Model	Accuracy, %	Training time	Params (M)
1	resnet50d	100	6m	25.6
2	vgg16	100	8m55s	138.36
3	vgg19	100	14m13s	143.67
4	convit_base	100	17m23s	86.54
5	crossvit_base	99.5	32m49s	105.03
6	vit_base_resnet50	98.51	31m42s	98.95
7	vit_base_patch16	92.04	13m56s	86.54

Table 4.1: Classification accuracy for CK+48 dataset with different models.

validation set, then report on its accuracy on the test set.

№	Model	Accuracy, %	Training time	Params (M)
1	resnet50d	92.86	2m4s	25.6
2	vgg16	92.86	3m37s	138.36
3	vgg19	94.29	2m	143.67
4	convit_base	91.42	3m22s	86.54
5	crossvit_base	90	3m28s	105.03
6	vit_base_resnet50	95.71	6m15s	98.95
7	vit_base_patch16	97.14	4m45s	86.54

Table 4.2: Classification accuracy for JAFFE dataset with different models.

Before delving into the specifics of how the models utilized perform on various datasets, we will go through our training approach quickly. We trained one model for each database in our trials, although we attempted to keep the structure and hyper-parameters consistent across models. Each model was trained on 50 epochs using the computing resources of Google Colab Pro. For optimization, we used stochastic gradient descents optimizer with a batch size of 10 and learning rate of 0.003 (Various values of the batch size and learning rate were tested, and the selected coefficients showed the best results). It took less than 10 minutes to train the pre-trained models on the JAFFE and CK+48 datasets, since the number of images is not so large, 213 and 981 images, respectively. However, it took a little longer to train some models of ViT, approximately 20-30 minutes. The images in the training sets are augmented with data to train the model on a greater number of images and make the learned model invariant on tiny modifications.

The CK+48 dataset exceeds JAFFE in the number of images. So, the number of training data is 781 and 143, respectively, while the tested images are 200 and 70 in each set. In turn, we would like to note some imbalance in the number of examples of emotions in the CK+48 dataset. For example, the emotion of surprise and happiness have 200 and 165 images, but the neutral emotion and the emotion of fear are represented by only 43 and 60 images, respectively.

The learning curve for the CK+48 and JAFFE datasets on various pre-trained models is shown in Figures 4-3 - 4-9. As we can observe from the experimental curves, the training of pre-trained models on the CK48 dataset is much faster than on the JAFFE dataset. Up to 10 epochs in the case of SK48 versus 25-30 epochs with JAFFE. The explanation for this can be the number of images in the data set, since CK48 exceeds JAFFE by about 5 times in volume. However, if we look at the training of ViT models, we will see that the training is faster. This behavior is explained by the fact that transformers require less resources compared to other models.

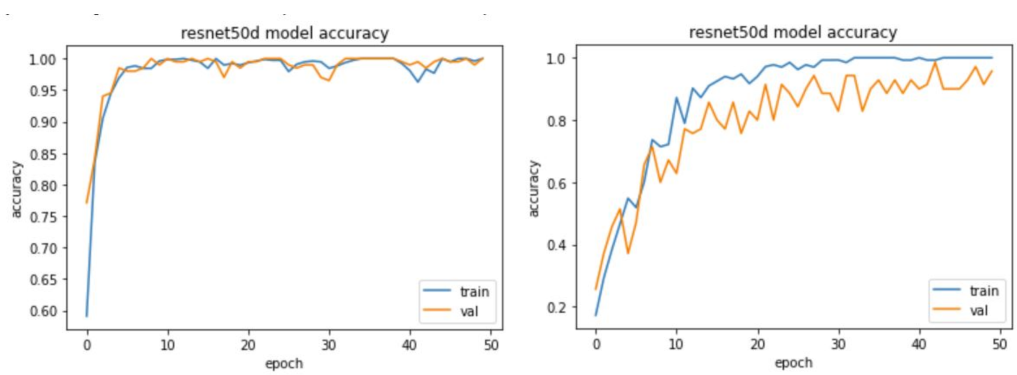

Figure 4-3: Training and Test accuracies of Resnet50d on CK+48 (left) and JAFFE(right) datasets

The confusion matrices on the test set of CK+48 and JAFFE datasets are shown in Figures4-10 - 4-13.

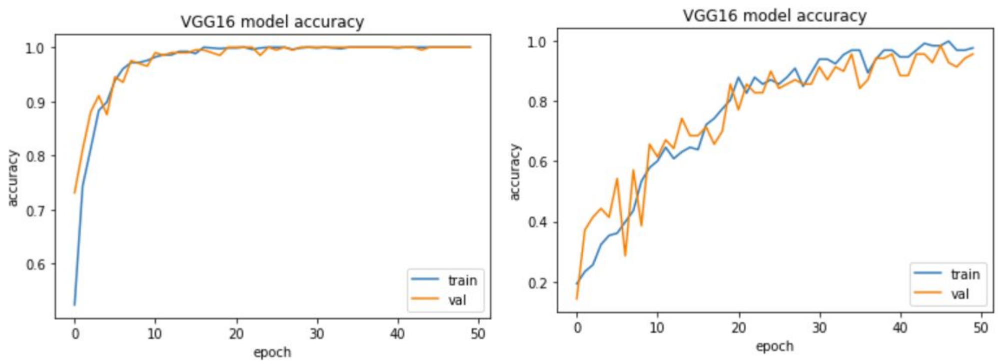

Figure 4-4: Training and Test accuracies of VGG16 on CK+48 (left) and JAFFE(right) datasets

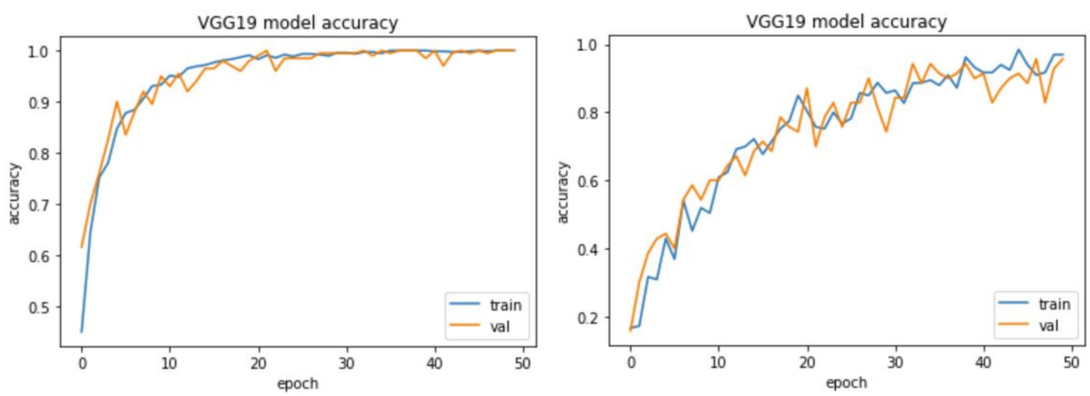

Figure 4-5: Training and Test accuracies of VGG19 on CK+48 (left) and JAFFE(right) datasets

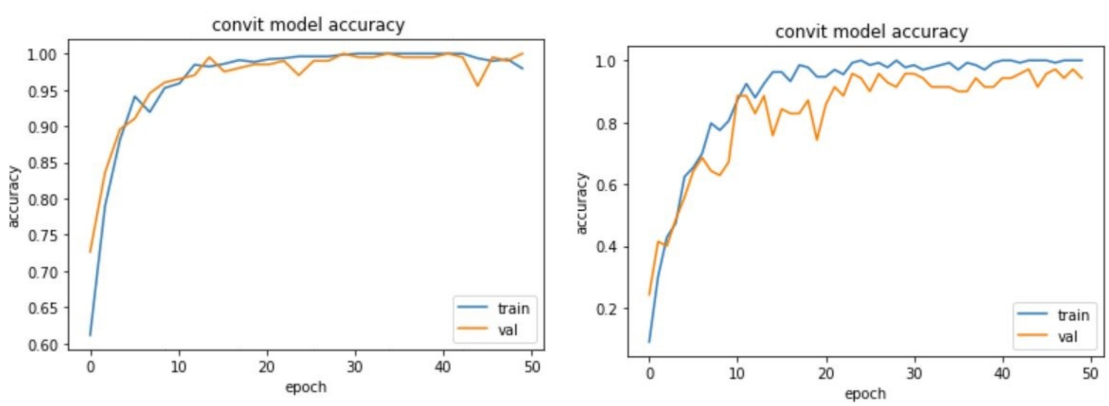

Figure 4-6: Training and Test accuracies of ConViT model on CK+48 (left) and JAFFE(right) datasets

4.4 Comparisons with state-of-the-arts

The proposed method is compared with other methods on the JAFFE database (Table 4.3). The JAFFE database is collected in a controlled laboratory environment, and

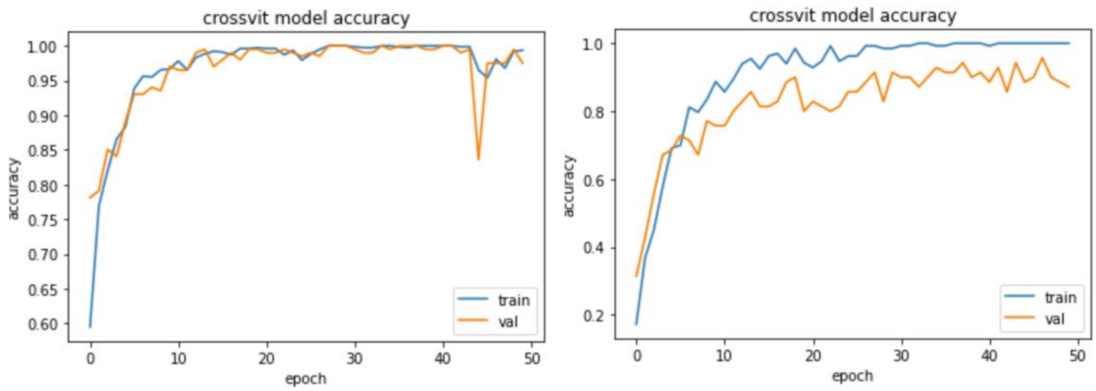

Figure 4-7: Training and Test accuracies of CrossViT model on CK+48 (left) and JAFFE(right) datasets

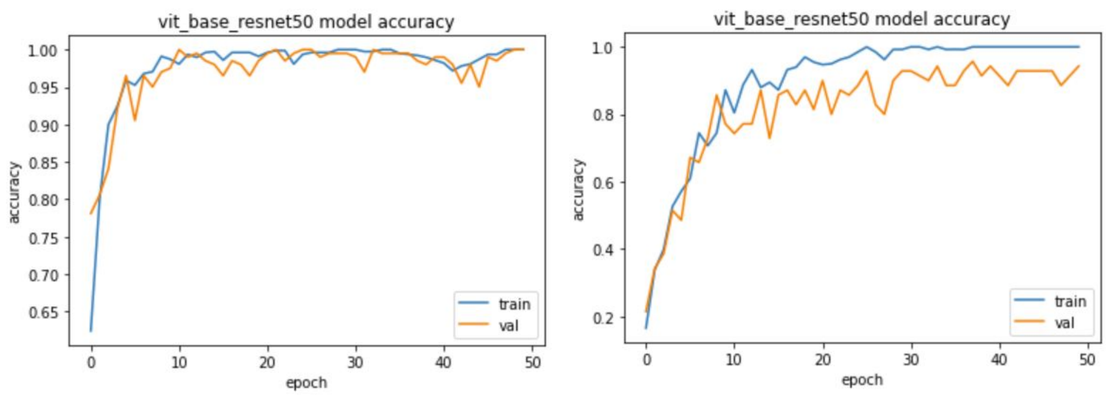

Figure 4-8: Training and Test accuracies of ViT_base_resnet50 model on CK+48 (left) and JAFFE(right) datasets

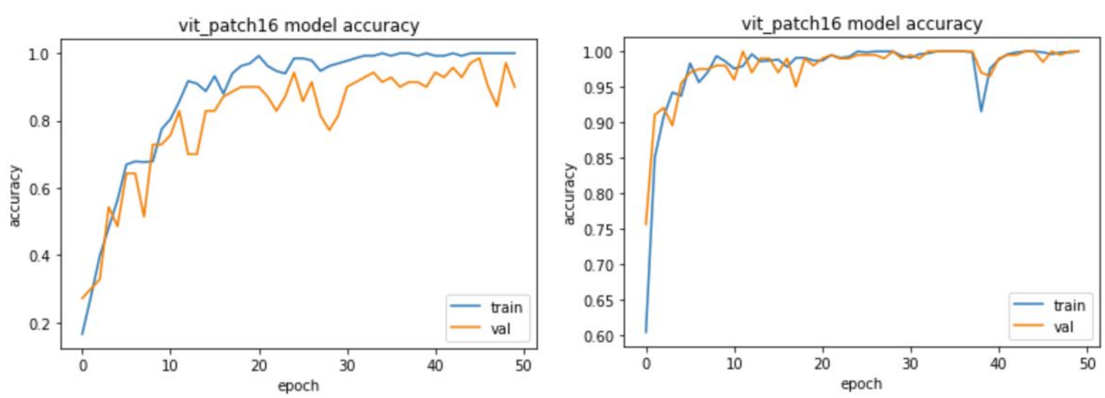

Figure 4-9: Training and Test accuracies of ViT_base_patch16 model on CK+48 (left) and JAFFE(right) datasets

all the data are frontal faces that have minor background changes. Table 4.3 shows that the proposed ViT models have an excellent validation accuracy on JAFFE with

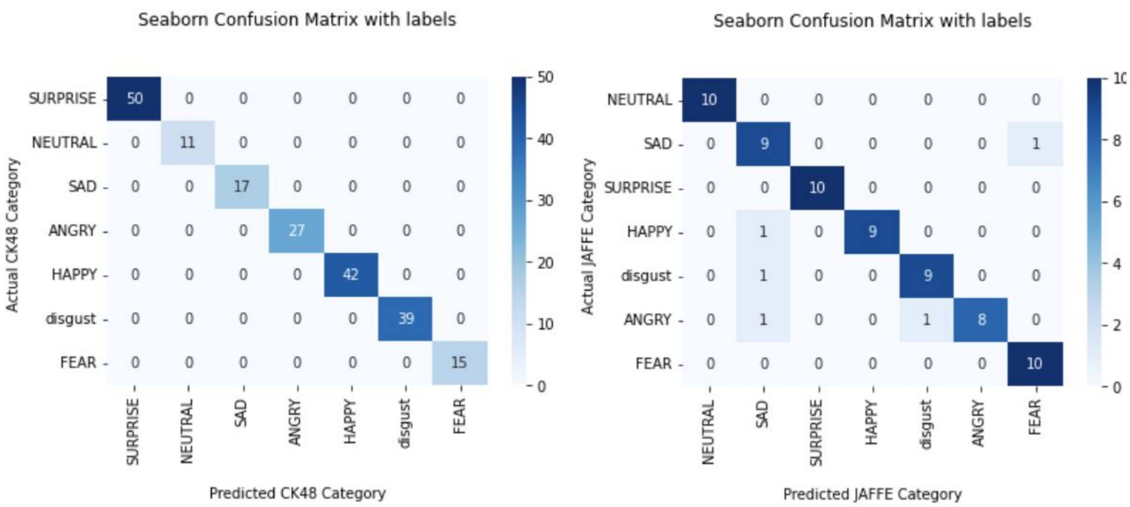

Figure 4-10: Confusion matrices of resnet50d model on CK+48 (left) and JAFFE(right) datasets

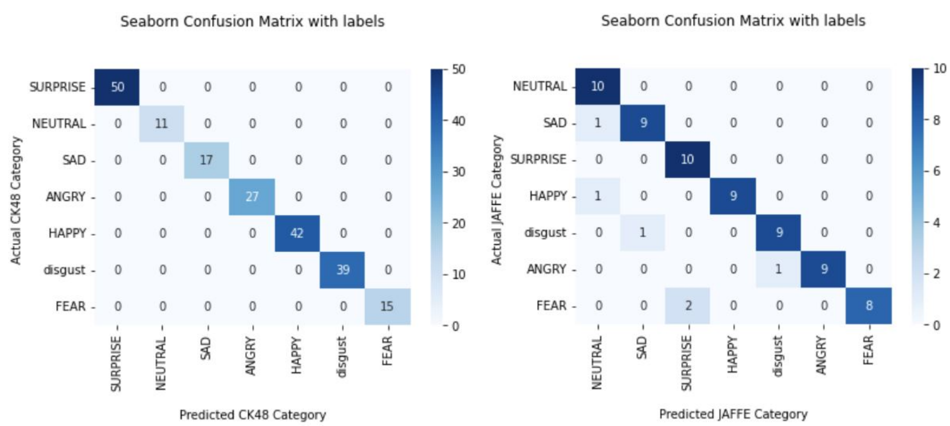

Figure 4-11: Confusion matrices of VGG19 model on CK+48 (left) and JAFFE(right) datasets

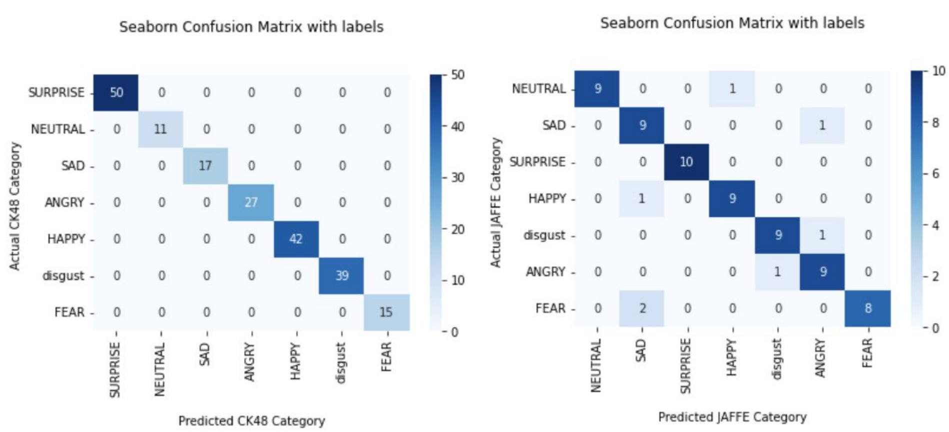

Figure 4-12: Confusion matrices of ConViT model on CK+48 (left) and JAFFE(right) datasets

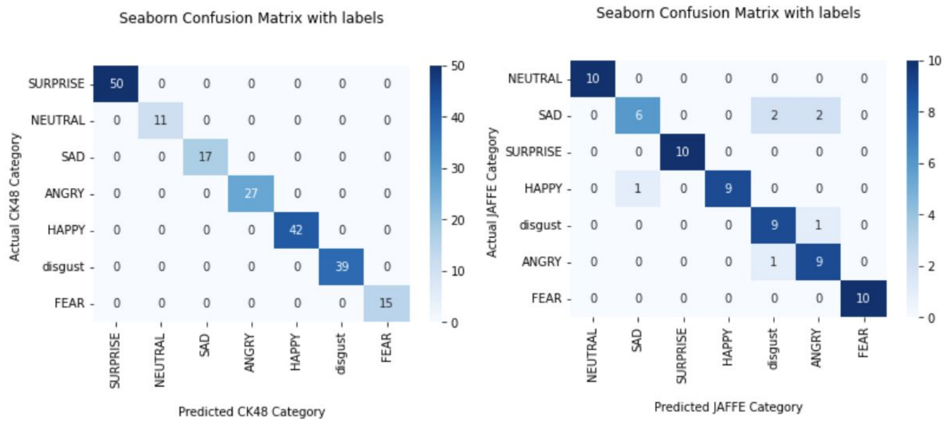

Figure 4-13: Confusion matrices of CrossViT model on CK+48 (left) and JAFFE(right) datasets

97.14%. As shown in Table 4.3, Mahesh et al. [30] use a a method of concatenating spatial pyramid Zernike moments based shape features which achieves an average accuracy of 95.86%. Boughida et al. [6] propose facial expression recognition approach based on Gabor filters and genetic algorithm and obtain an accuracy of 96.3%. Proposed by Liu et al. [28] method focuses more on the facial feature extraction on the basis of facial landmarks, helping the network extract more discriminative features that are conducive to recognize expressions. Their method uses a Spatial Attention Convolutional Neural Network (SACNN) to extract the pixel-level facial feature and employs Long Short-term Memory networks with Attention mechanism (ALSTMs) to explore the deep geometric position correlation of facial landmarks. The facial landmarks are divided into seven groups for local-holistic geometric feature extraction and the attention mechanism is utilized to estimate the importance of different landmark regions. Thus, the combination of the attention mechanism together with the geometric correlation of the positions of facial landmarks gives the best recognition result - 98.57%, which exceeds our method by about 1.5%.

The advantage of ViT models based on the attention mechanism is reduced training time and consumption of less computing resources.

Approach	Accuracy (%)
Mahesh et al. [30]	95.86
Boughida et al. [6]	96.3
Liu et al. [28]	**98.57**
Aouayeb et al. [1]	92.92
Our (vit_base_resnet50)	95.71
Our (vit_base_patch16)	**97.14**

Table 4.3: Comparison with state-of-the-art methods on JAFFE

4.5 Model visualization and analysis

Here we propose an approach to visualizing classification accuracy using a dimensionality reduction technique t-SNE.

The t-SNE - is an algorithm for dimensionality reduction. This algorithm allows us to visualize the high-dimensional data of the facial images. The t-SNE function will convert high dimensional data into low dimensional data. Generally, distant points in high-dimensional space will be converted into distant embedded low-dimensional points and nearby points in the high-dimensional space will be converted into nearby embedded low-dimensional points. As a result, we can visualize the low-dimensional points to find the clusters in the original high-dimensional data

Figure 4-14 shows the t-SNE 3D plot of the extracted features form the vit_base_patch16 model on JAFFE dataset. As we can see from the graph, the features extracted from the JAFFE dataset show similar values, as a result of which the visualization of the division into classes turned out to be not clear, where all 7 categories of facial expressions are mixed.

Figure 4-15 shows the t-SNE 3D plot corresponding to the 768-dimensional features from the ViT model. The features correspond to the CK+48 images. In the case of the CK+48 dataset, which exceeds the number of JAFFE images by approximately 5 times, data visualization shows slight improvements. So we see that such classes as happiness (purple) and surprise (pink) are clearly grouped on both sides of the graph. In the central part, the remaining 5 classes are mixed.

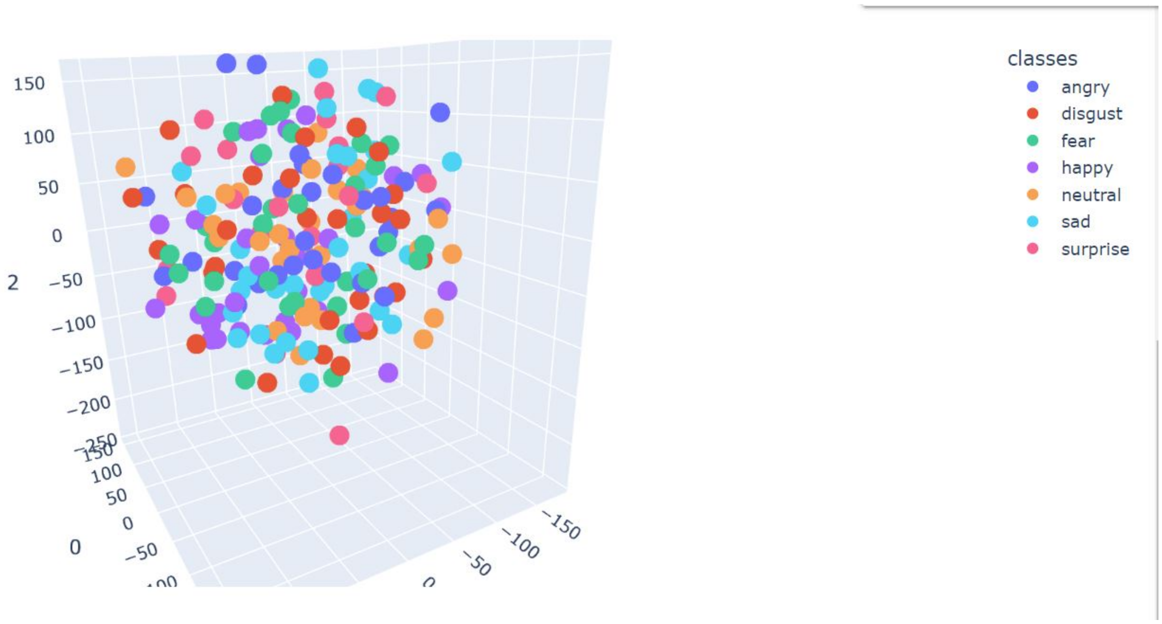

Figure 4-14: t-SNE 3D plot on JAFFE dataset

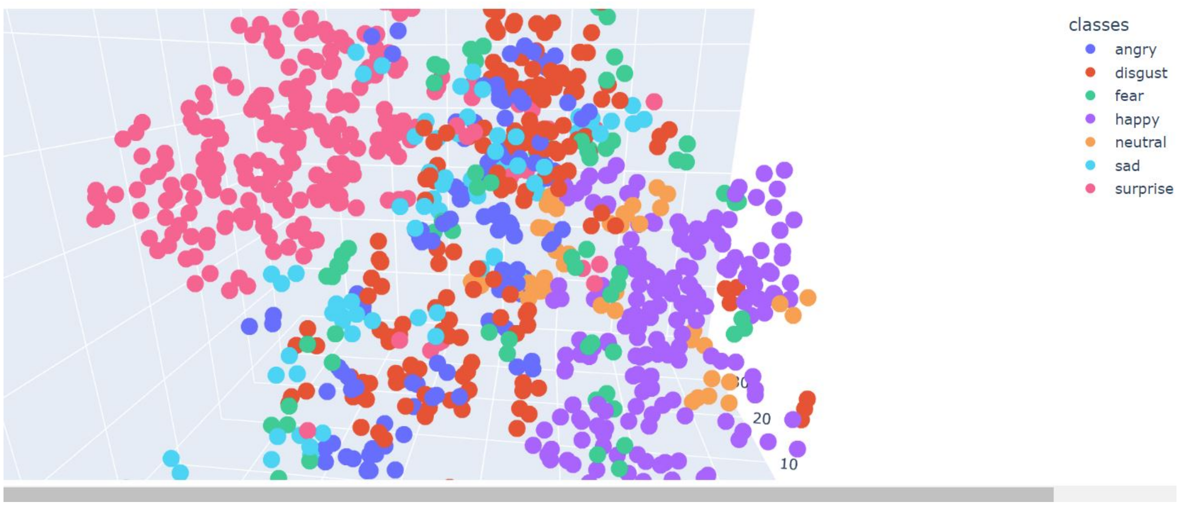

Figure 4-15: t-SNE 3D plot on CK+48 dataset

Chapter 5

Conclusion and Future directions

The attention mechanism can direct the network's attention to critical feature information while suppressing background disturbance. Because of its basic structure, low complexity, and few parameters, the network in this paper can train and forecast the model quickly and effectively. Compared to CNN, the ViT model uses multi-head self-control without requiring image-specific biases. In addition, ViT has a higher precision rate for a large dataset with reduced training time. We have presented the classification results on lab-made databases (CK+48 and JAFFE) to evaluate the performance of the selected models. The main contribution of our research is to conduct empirical studies with ViT models that have achieved relatively good results on such publicly available datasets with modern facial expression recognition algorithms. The results imply that representations gained from pre-trained networks taught for a specific task, such as object detection, can be transferred and used for a different task, such as facial expression recognition.

We are optimistic about the future of attention-based models and intend to apply them to other activities. Our further research will be the use of a ViT models not only on static images, but also on a sequence of such images taking into account the time parameter (for example, original CK+ and Oulu-CASIA datasets). Another way of further research is the use of ViT models on larger datasets obtained in the wild conditions, such as FER-2013, SFEW and RAF-DB.

Bibliography

[1] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

[4] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. *arXiv preprint arXiv:2102.08602*, 2021.

[5] Manoj Bhat, Jonathan Francis, and Jean Oh. Trajformer: Trajectory prediction with local self-attentive contexts for autonomous driving. *arXiv preprint arXiv:2011.14910*, 2020.

[6] Adil Boughida, Mohamed Nadjib Kouahla, and Yacine Lafifi. A novel approach for facial expression recognition based on gabor filters and genetic algorithm. *Evolving Systems*, 13(2):331–345, 2022.

[7] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[8] Junkai Chen, Zenghai Chen, Zheru Chi, Hong Fu, et al. Facial expression recognition based on facial components detection and hog features. In *International workshops on electrical and computer engineering subfields*, pages 884–888, 2014.

[9] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021.

[10] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019.

[11] Arnaud Dapogny and Kevin Bailly. Investigating deep neural forests for facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 629–633. IEEE, 2018.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

[16] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[17] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[19] Yingruo Fan, Victor Li, and Jacqueline CK Lam. Facial expression recognition with deeply-supervised attention network. *IEEE transactions on affective computing*, 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[22] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2017.

[23] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *arXiv preprint arXiv:1702.00887*, 2017.

[24] Jing Li, Kan Jin, Dalin Zhou, Naoyuki Kubota, and Zhaojie Ju. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350, 2020.

[25] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

[26] Wei Li, Min Li, Zhong Su, and Zhigang Zhu. A deep-learning approach to facial expression recognition with candid images. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 279–282. IEEE, 2015.

[27] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2209–2214. IEEE, 2018.

[28] Chang Liu, Kaoru Hirota, Junjie Ma, Zhiyang Jia, and Yaping Dai. Facial expression recognition using hybrid features of pixel and geometry. *Ieee Access*, 9:18876–18889, 2021.

[29] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Shilei Sun, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020.

[30] Vijayalakshmi GV Mahesh, Chengji Chen, Vijayarajan Rajangam, Alex Noel Joseph Raj, and Palani Thanaraj Krishnan. Shape and texture aware facial expression recognition using spatial pyramid zernike moments and law's textures feature set. *IEEE Access*, 9:52509–52522, 2021.

[31] Albert Mehrabian. Communication without words. pages 193–200, 2017.

[32] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[33] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[34] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Aravind Ravi. Pre-trained convolutional neural network features for facial expression recognition. *arXiv preprint arXiv:1812.06387*, 2018.

[36] Aravind Ravi, Harshwin Venugopal, Sruthy Paul, and Hamid R Tizhoosh. A dataset and preliminary results for umpire pose detection using svm classification of deep features. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1396–1402. IEEE, 2018.

[37] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[38] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[39] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[40] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021.

[43] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[44] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[47] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Supplementary material for ‘eca-net: Efficient channel attention for deep convolutional neural networks. Technical report, Tech. Rep.

[48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[49] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021.

[50] Jacob Whitehill and Christian W Omlin. Haar features for facs au recognition. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 5–pp. IEEE, 2006.

[51] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[52] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[53] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv e-prints*, pages arXiv–2012, 2020.

[54] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[55] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442, 2015.

[56] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.